Les 6. De χ^2 toets.

Onderzoekstechnieken

Jens Buysse **Wim De Bruyn** **Bert Van Vreckem**

AJ 2018-2019

HO GENT

What's on the menu today?

- χ^2 toets voor één variabele
  - Toetsingsprocedure goodness of fit test
  - Voorbeeld
  - Gestandaardiseerde residuen
- χ^2 toets voor twee variabelen

HO GENT

Wat weten we nog van vorige les?

- Wat is een hypothese
- Wat zijn de onderdelen van een hypothesetoets
- Wat zijn de stappen bij het toetsen?
- Welke fouten kunnen er gemaakt worden?

HO GENT

χ^2 toets voor één variabele

HO GENT

Goodness of fit test

De goodness of fit test kan gebruikt worden om na te gaan in welke mate de steekproef overeenstemt met een nulhypothese over de verdeling van de variabele.

HO GENT

Goodness of fit test

We willen nagaan of de verdeling van onze steekproef bij $n = 400$ superhelden overeenstemt met de verdeling die je verwacht in de volledige populatie (de verzameling van alle mogelijke superhelden).

- We Vergelijken daartoe de aantallen in de steekproef met de aantallen die je zou verwachten als de steeproef exact representatief zou zijn naar de types van superhelden.
- Verschillen relatief groot $\Rightarrow$ dan komt de verdeling in de steekproef niet overeen
- Verschillen relatief klein $\Rightarrow$ dan komt de verdeling in de steekproef overeen

HO GENT

Goodness of fit test

We willen nagaan of de verdeling van onze steekproef bij $n = 400$ superhelden overeenstemt met de verdeling die je verwacht in de volledige populatie (de verzameling van alle mogelijke superhelden).

- We Vergelijken daartoe de aantallen in de steekproef met de aantallen die je zou verwachten als de steeproef exact representatief zou zijn naar de types van superhelden.
- Verschillen relatief groot $\Rightarrow$ dan komt de verdeling in de steekproef niet overeen
- Verschillen relatief klein $\Rightarrow$ dan komt de verdeling in de steekproef overeen

Zie je een overeenkomst bij kruistabellen en Cramer's V?

HO GENT

Goodness of fit test

Type superheld	# in steekproef	# in populatie
Mutant	127	35%
Mens	75	17%
Alien	98	23%
God	27	8%
Demon	73	17%

HO GENT

Goodness of fit test

- We willen kijken of de steekproef voor onze superhelden representatief is.
- Als de steekproef exact representatief is $\Rightarrow$ in steekproef 35% van de superhelden een mutant
- Het verwachte aantal is dus gelijk aan $0.35 \times 400 = 140$.
- De verwachte frequenties worden genoteerd met de letter ***e*** (expected).

Er geldt dus:

$$e = n \times \pi$$

Als de verschillen $o - e$ relatief klein zijn kunnen ze toegerekend worden aan toevallige steekproeffouten.

HO GENT

Goodness of fit test

Beschouw χ^2:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

We merken op:

- indien de verschillen klein zijn $\Rightarrow$ verdeling komt voldoende overeen
- indien de verschillen groot $\Rightarrow$ verdeling niet representatief

χ^2 meet de mate van strijgdigheid van de gegevens met de nullhypothese

HO GENT

Goodness of fit test

Type superheld	o	π	e	$o - e$	$\frac{(o-e)^2}{e}$
Mutant	127	35%	140	-13	1.21
Mens	75	17%	68	7	0.72
Alien	98	23%	92	6	0.39
God	27	8%	32	-5	0.78
Demon	73	17%	68	5	0.37

HO GENT

Goodness of fit test

- De teststatistiek χ^2 is verdeeld volgens de chi kwadraat verdeling.
- Kritieke grenswaarde g bij de χ^2 verdeling: hierbij speel het aantal vrijheidsgraden een rol (df). Er geldt:

$$df = k - 1$$

  met k het aantal categorieën.
- $df = 5 - 1 = 4$.
- De kritieke waarden voor een gegeven significantieniveau α en vrijheidsgraden df is opnieuw gegeven in tabel.

In ons voorbeeld is $\chi^2 = 3.47$ met grenswaarde $g = 9.49$ en besluiten we omdat $3.47 < 9.49$ dat de steekproef representatief is (H_0 wordt aanvaard).

HO GENT

Toetsingsprocedure goodness of fit test

1. **Bepalen hypotheses**
   - H_0: steekproef is representatief naar populatie
   - H_1: steekproef is niet representatief naar populatie
2. **Bepalen α en n** : $\alpha = 0.05$ en $n = 400$.
3. **Toetsingsgrootheid en waarde ervan in steekproef**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

4. **Bereken en teken kritiek gebied**: de toets is altijd rechtszijdig. Is de toetsingsgrootheid kleiner dan kritieke grenswaarde verwerp H_0 niet, anders verwerp H_0 en aanvaard H_1.

HO GENT

Voorbeeld gezinnen

Beschouw alle gezinnen met 5 kinderen in een bepaalde gemeenschap.

HO
GENT

Voorbeeld gezinnen

Beschouw alle gezinnen met 5 kinderen in een bepaalde gemeenschap.
Met betrekking tot samenstelling zijn er 6 mogelijkheden.

1. 5 jongens
2. 4 jongens, 1 meisje
3. 3 jongens, 2 meisjes
4. 2 jongens, 3 meisjes
5. 1 jongen, 4 meisjes
6. 5 meisjes

Het onderzoek bevat 1022 gezinnen met 5 kinderen

Zijn de waargenomen aantallen in de 6 klassen representatief voor een populatie waar de kans om een jongen te krijgen = kans om een meisje te krijgen = 0.5?

HO GENT

Voorbeeld

i	0	1	2	3	4	5
o_i	58	149	305	303	162	45

HO GENT

Voorbeeld

i	0	1	2	3	4	5
o_i	58	149	305	303	162	45

Indien de veronderstelling waar is wordt de kans π_i om i jongens te krijgen bepaald door een binominaalvedeling met parameters $n = 5$ en $p = 0.5$. Bv. De kans om 2 jongens te krijgen met 5 kinderen is gelijk aan :

$$(0.5)^2 \times (1 - 0.5)^{5-2} \times \binom{5}{2}$$

Algemeen geldt dus:

$$\pi_i = \binom{5}{i} \times 0.5^i \times 0.5^{5-i} = \frac{5!}{i!(5-i)!} \times 0.5^i$$

HO GENT

Voorbeeld

i	0	1	2	3	4	5	
o_i	58	149	305	303	162	45	1022
π_i	0.03	0.15	0.31	0.31	0.15	0.031	1
e_i	31.68	159.43	318.86	318.86	159.43	31.68	
$\frac{(o-e)^2}{e}$	21.86	0.68259	0.60	0.78	0.041	5.59	29.57
r_i	4.74	-0.89	-0.93	-1.07106	0.22	2.40	

HO GENT

Voorbeeld

1. **Bepalen hypotheses**
   - H_0: steekproef is representatief naar populatie
   - H_1: steekproef is niet representatief naar populatie
2. **Bepalen α en n** : $\alpha = 0.01$ en $n = 1022$.
3. **Toetsingsgrootheid en waarde ervan in steekproef**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} = 29.5766$$

4. **Bereken en teken kritiek gebied**: kritieke grens is 15.0863. Onze toetsingsgrootheid ligt dus in het kritieke gebied dus verwerpen we H_0.

HO GENT

Gestandaardiseerde residuen

De gestandaardiseerde residuen duiden aan welke klassen de grootste bijdrage leveren aan de waarde van de grootheid.

$$r_i = \frac{O_i - n\pi_i}{\sqrt{n\pi_i(1-\pi_i)}}$$

- Er geldt algemeen: waarden groter dan 2 of kleiner dan -2 zijn extreem.

We kunnen dus besluiten dat het aantal gezinnen waarin alle kinderen hetzelfde geslacht hebben groter mag worden genoemd dan verwacht.

HO GENT

Voowaarden

Om de toets te mogen toepassen dient aan de volgende voorwaarden te zijn voldaan (Regel van Cochran)

1. Voor alle categorieën moet gelden dat de verwachte waarde e groter is dan 1.
2. In ten hoogste 20 % van de categoriën mag de verwachte waarde e kleiner dan 5 zijn.

HO GENT

χ^2 toets voor twee variabelen

HO GENT

χ^2 toets voor twee variabelen

De Chi-kwadraattoets laat zich eenvoudig uitbreiden tot een onderzoeksontwerp met twee variabelen, met respectievelijk ***r*** en ***k*** niveaus.

HO GENT

Rokersonderzoek

In deze studie onderzochten Doll en Hill de relatie tussen roken en longkanker. Doll en Hill schreven in 1951 alle Britse huisartsen aan met het verzoek om gegevens over hun leeftijd en rookgedrag. Vervolgens hielden ze jarenlang de overlijdensberichten en de doodsoorzaak bij en herhaalden dit periodiek. De eerste uitkomsten, na circa vier jaar, zijn in de volgende tabel samengevat.

	Longkanker	Niet	Wel	Totaal
Roker	**Wel**	21178	83	21261
	Niet	3092	1	3093
	Totaal	24270	84	24354

HO GENT

Rokersonderzoek

	Longkanker	Niet	Wel	Totaal
Roker	Wel	21178	83	21261
	Niet	3092	1	3093
	Totaal	24270	84	24354

- ...slechts $\frac{84}{24354} \times 100 = 0.35\%$ van de Britse artsen aan longkanker overleden
- ...met slechts $\frac{83}{21261} \times 100 = 0.39\%$ van de rokers onder hen
- ...maar is wel meer dan hetzelfde cijfer voor de niet-rokers $\frac{1}{3093} * 100 = 0.032\%$.

HO GENT

Rokersonderzoek

	Longkanker	Niet	Wel	Totaal
Roker	Wel	21188	73.3	21261
	Niet	3082.3	10.7	3093
	Totaal	24270	84	24354

- $\chi^2 = 10.35$
- We zien in de tabel dat er wel een erg groot verschil is tussen de geobserveerde aantallen rokers die overlijden aan longkanker en de verwachte waarden in deze cel.
- Hetzelfde geldt voor het geringe aantal huisartsen dat niet rookt, maar wel aan longkanker overleden is.

Rokersonderzoek

1. **Bepalen hypotheses**
   - H_0: in de populatie is er geen samenhang tussen onafhankelijke en afhankelijke variabele
   - H_1: er bestaat wel een samenhang tussen de variabelen in de populatie
2. **Bepalen α en n** : $\alpha = 0.05$ en $n = 24354$.
3. **Toetsingsgrootheid en waarde ervan in steekproef**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{E_i} = 10.35$$

4. **Bereken en teken kritiek gebied**: kritieke grens is 3.8415 en aantal vrijheidsgraden $df = (r-1)(k-1)$ Onze toetsingsgrootheid ligt dus in het kritieke gebied dus verwerpen we H_0.

HO GENT

Oorzakelijk verband

We moeten derhalve H_0, dat er geen relatie is tussen beide variabelen, verwerpen ten gunste van H_1 dat er wel een relatie is tussen beide variabelen: rokers sterven vaker aan longkanker dan niet-rokers.

- ...rokers zijn ouder dan de niet-rokers
- ...de rokers wonen veelal in de grote steden met meer vervuilde lucht dan de niet-rokers
- ...speciale genetische dispositie die zowel van invloed is op de verslaving aan tabak, als op de kans om longkanker te krijgen.

Voor een causale interpretatie van de gegevens (het betreft hier immers geen experiment), moeten we op zijn minst de beschikking hebben over een theorie die de relatie tussen roken en longkanker expliciteert.

HO GENT